Robust real-time hand detection and localization for space human-robot interaction based on deep learning

Qing Gao, Jinguo Liu, Zhaojie Ju

Abstract

Hand gestures are quite suitable for space human-robot interaction (SHRI) because of their natural and convenient features. While the detection and localization of hands are the premise and foundation for SHRI based on hand gestures. But hand gestures are very complicated and hand sizes are very small in some images. These problems make the robust real-time hand detection and localization very difficult. In this paper, a feature-map-fused single shot multibox detector (FF-SSD) which is a deep learning network is designed to deal with the problems of hand detection and localization in SHRI. First, the background of the method is introduced in this paper, including an astronaut assistant robot platform, the difficulties of hand detection and localization, and introduction of the state-of-the-art deep learning networks for object detection and localization. Then, the FF-SSD is proposed for detecting and localizing hands especially pony-size hands. This network magentatakes into consideration both accuracy and speed with balanced performance. And in the experiment part, the FF-SSD is trained and tested on hand databases which include a homemade database and two public databases. At last, the superiority of the proposed method is demonstrated compared with the state-of-the-art methods.

Keywords: Astronaut assistant robot, Deep learning, Hand detection and localization, SSD

2018 MSC: 00-01, 99-00

March 10, 2019

1. Introduction

Space robots are required to assist astronauts in some tasks in the space station because of the limited number of astronauts and the heavy space tasks. In-cabin flying robots can fly and hover in the cabin, and most of them have small sizes. So, they are quite suitable for assisting astronauts in the cabin. Astronauts can interact with the robots face-to-face via hand gestures or voice. While the robots can collect information and send it to astronauts to monitor in-cabin environment and equipment.

Nowadays, there are many famous in-cabin flying robots, some of them have even been tested in the International Space Station (ISS). Such as Personal Satellite Assistant (PSA) [1], this robot is developed by NASA AMS Research Center. It is employed to analyzing and monitoring in-cabin environment and equipment, and it also can be used for health management for astronauts. Smart SPHERES [2] is designed by MIT. Astronauts can supervise the cabin environment by controlling this robot remotely. Astrobee [3, 4] is inherited from PSA. It's an in-cabin flying robot developed by NASA in the Human Exploration Telerobotics2 (HET2) project. Crew interactive mobile companion (CIMON) [5] is a special robot which is customized for astronauts by NASA and IBM. It can freely float and move in the ISS and communicate with astronauts and assist astronauts in some work. Take into account these robots, we also design an in-cabin flying robot named astronaut assistant robot (AAR) [6, 7]. Astronauts can use hand gestures to communicate with this robot face-to-face [8].

In the process of SHRI based on hand gestures [9, 10, 11, 12], the detection and localization of the astronaut's hands are of great importance. They are the premise and basis of gesture recognition and hand tracking. However, it is difficult to detect and locate astronauts' hands. There are several reasons. First of all, there are numerous hand gestures for SHRI. Second, it needs strong real-time capability. What's more, when the distance between the astronaut and the robot is long enough, the image size of hand is small and the resolution of hand image is very low. These situations increase the difficulties of hand

detection and localization.

At present, deep learning methods have achieved good results in the field of object detection and localization. Typical deep learning models for object detection and localization are Region-based Convolutional Neural Networks (R-CNN) [13], Fast-RCNN [14], Faster-RCNN [15], You Only Look Once (YOLO) [16] and Single Shot MultiBox Detector (SSD) [17]. Among them, SSD is a very effective model, it takes into consideration the object detection accuracy and real-time capability. Its accuracy is higher than YOLO and speed is faster than Faster-RCNN. However, because of the characteristics of the network structure of SSD, the detection of small objects is not very good. SSD model is not a good choice when a hand image is regarded as a small object. Therefore, it is necessary to improve its framework to improve the detection accuracy for small objects. Models such as Deconvolutional Single Shot Detector (DSSD) [18, 19, 20] have improved the framework of SSD. But balancing accuracy and real-time is still a challenge.

This paper, focusing on the SHRI based on hand gestures, mainly studies the astronaut's hand detection and localization method using deep learning networks. The main contributions of our work are summarised as follows:

- The proposed FF-SSD can deal with the detection of complex hand gestures and pony-size hand images with a high accuracy.
- The proposed network can achieve a satisfactory efficiency while ensuring a high detection accuracy.
- The proposed method has been successfully implemented into the astronaut assistant robot.

The rest of this paper is organized as follows: in chapter 2, backgrounds of FF-SSD are introduced, they include the AAR platform, the difficulties of hand detection and localization, and the introduction of state-of-the-art networks for object detection and localization. In chapter 3, the FF-SSD network is

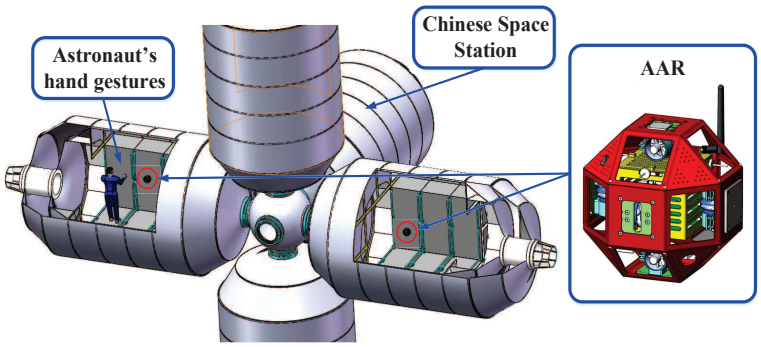

Figure 1: The imaginary diagram that an astronaut uses hand gestures to control the AAR.

proposed. In chapter 4, the comparative experiment results are shown to prove the superiority of the FF-SSD.

2. Background

2.1. AAR Platform

The AAR is built for assisting astronauts to finish some space tasks. In order to make astronauts communicate with the robot face-to-face, a hand-gesture-based SHRI method is designed to make the communication naturally. Astronauts can use hand gestures to control the AAR when they are doing some other things. Figure 1 is an imaginary diagram that an astronaut uses hand gestures to interact with the AAR in the space station.

For the hand gestures of astronauts, a set of natural and reasonable hand gestures is required for the hand gestures interaction between astronauts and AAR. American Signal Language (ASL) hand dataset [21] contains 26 kinds of hand gestures, representing 26 English letters. They are easy to learn and understand, so it is very suitable for SHRI. As a result, 8 hand gestures are selected from ASL for using in SHRI, and the designed SHRI hand dataset is shown in Table 1.

For the AAR platform, it includes the AAR, an image capturing device (Kinect v2) and an air floatation simulator. The robot uses the Kinect v2

Table 1: SHRI hand gestures dataset.

ASL meaning	Semantic meaning	Hand gestures
B	Begin to control	
S	Stop control	
F	Finish an action	
P	Path tracking	
L	Linear motion	
R	Rotational motion	
O	Object approaching	
D	Data transmission	

to collect astronaut's hand images, and then utilizes the hand detection and localization system to locate hands to prepare for subsequent hand gestures recognition and hand tracking. Where the air floatation simulator can simulate the microgravity environment in space station. Kinect v2 can collect RGB images and deep images of hands. Onboard computer mainly includes GTX 1060 and STM32. From Figure 2, we can see that an astronaut communicates with the robot face-to-face through hand gestures. And the first step of astronaut-AAR interaction is detection and localization astronaut's hand.

2.2. Hand detection and localization

It is very difficult to detect and locate the gestures in Table 1. First of all, unlike the detection of other objects such as vehicles, the changes between different hand gestures are very large. This problem greatly increases the difficulty of hand detection. In addition, human hand is a non-rigid object with a complex structure. And different people's hands have different sizes, shapes and colors.

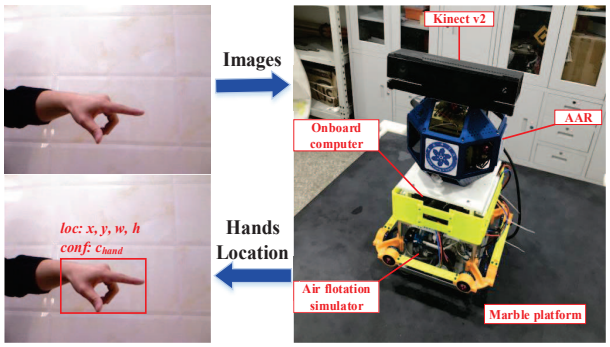

Figure 2: SHRI platform.

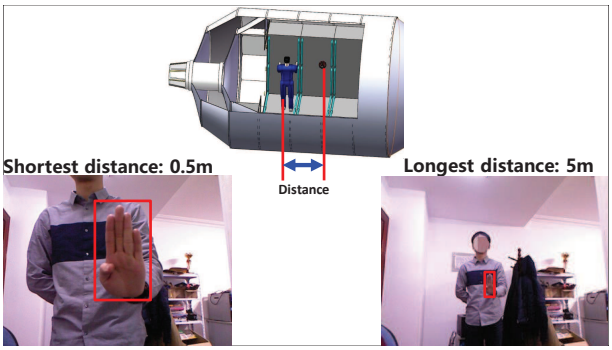

Figure 3: Hand image sizes under different distances.

In addition, before performing hand gestures interaction, the distance between astronaut and robot should be considered. On the one hand, considering the safety of astronauts, the AAR robot should keep a certain distance from the astronauts to avoid collisions. And the minimum distance is set to 0.5 meters. On the other hand, considering the limited space in the cabin of the space station, there should be a maximum distance and this distance is set to 5 meters. And Figure 3 shows hand image sizes under the maximum distance and minimum distance.

As can be observed in Figure 3, when the distance between the astronaut and the AAR robot is the maximum distance, the size of the hand image is very small. The smaller hand size is, the lower the hand image resolution is. It makes the hand detection and localization more difficult.

In summary, the designed astronauts hand detection and localization method should meet the following conditions.

- It can detect and locate hands in real time.
- It can detect all kinds of SHRI hand gestures.
- It can get a high accuracy of hand detection and localization when hand image size is very small.

2.3. The state-of-the-art networks

Among the deep learning models for object detection and localization, the SSD [17] has a very good performance, which not only has a high accuracy, but also has a fast speed for real-time detection and localization. Its structure is shown in Figure 4(a). It is a detector based on a full convolutional network that uses different layers to detect objects of different sizes. However, the SSD model cannot get a good performance to small targets. Because the shallower layers of the network have big sized feature maps which have enough contextual information but less semantic information. While its deeper layers have sufficient semantic information but through too many pooling layers, the sizes of feature maps are too small. When the detected hand has a small image size, it needs both large feature map to provide enough detailed features, sufficient intensive sampling, and sufficient semantic meaning to distinguish hand from the background.

Aiming at the shortcomings of SSD's poor detection to small targets, some improved models have emerged [18, 19, 20, 22]. Among them, one of the most famous methods is DSSD [18]. It's illustrated in Figure 4(b). Firstly, it replaces the reference network of SSD from VGG-16 to Resnet-101. This can enhance the feature extraction capability. And then it uses the deconvolution layer to add a lot of contextual information. After deconvolution the feature map will have a higher resolution and richer contextual information, which can increase the detection accuracy of small targets. But the detection speed will be reduced. As the anterior part of hand gestures recognition, hand detection and localization

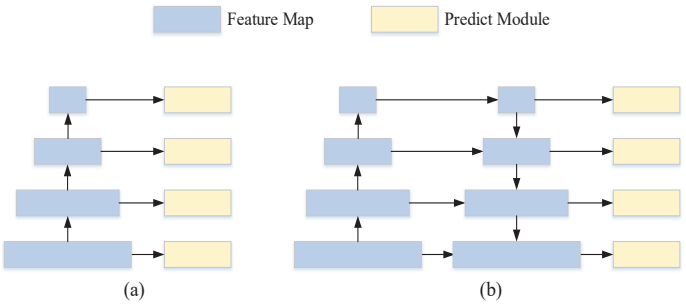

Figure 4: (a) is the SSD framework, (b) is the DSSD framework.

need a fast speed to satisfy the real-time SHRI system. So, the DSSD is not a good choice.

3. Feature-map-fused SSD

3.1. The proposed FF-SSD

Borrow the DSSD's idea that using deconvolution to combine deep layers with shallow layers to increase the amount of contextual information, the network can be simplified to increase the speed of detection and also ensure the detection accuracy of small objects.

According to the above, we change two parts for SSD network to improve its framework.

- *Replace VGG-16 layers with Resnet-101 layers.* As we know from reference [23], Resnet-101 can extract better features than VGG-16. In addition, compared with other networks, the structure of resnet101 is not that complicated, which can guarantee the real-time performance of the network. Because it uses shortcut connection framework to solve the performance degradation problem when the depth of network is increased. The building block of residual learning is illustrated in Figure 5. And the residual mapping formula is shown as :

$$y = F(x) + x \tag{1}$$

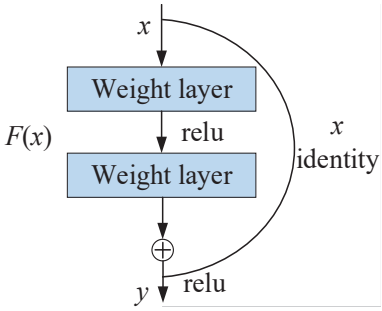

Figure 5: A building block of residual learning.

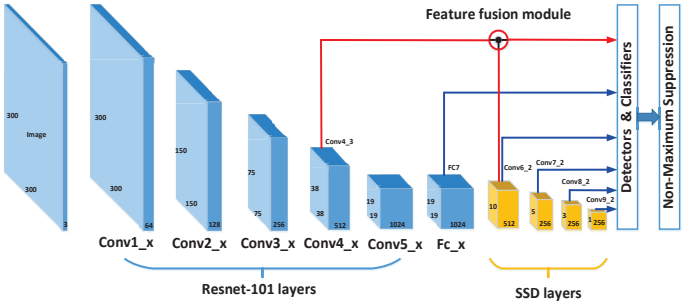

Figure 6: FF-SSD architecture.

where x indicates unit mapping, y indicates optimal solution, and $F(x)$ indicates the residual between the optimal solution and the unit map.

- *Fuse the features of Conv*4_3 *layer and Conv*6_2 *layer.* Because the SSD network uses its shallow layers to detect small objects, it can find out which shallow layer is most suitable for small targets detection by visualizing the shallow layers of the SSD network. As can be known from paper [22], when object image size is small, compared with Conv4_3 layer, Conv3_3 layer has insufficient effective receptive fields, and Conv5_3 layer and Fc6 layer have larger effective receptive fields, but they will introduce more background noise. Therefore, Conv4_3 layer is most suitable for the detection of small objects.

The improved network structure is shown in Figure 6.

As shown in Figure 6, the features of Conv6_2 layer and Conv4_3 layer are

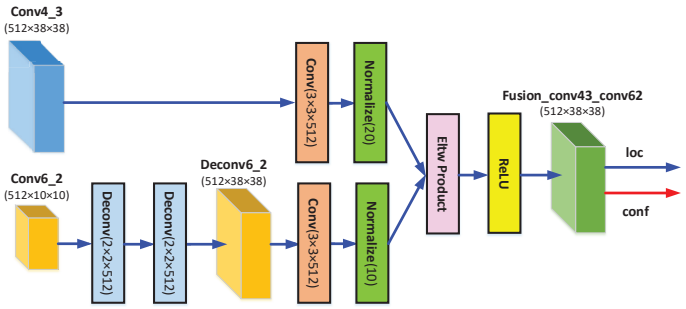

Figure 7: Feature fusion module.

merged to increase the semantic information to the Conv4_3 layer. Thereby the detection accuracy of small targets can be increased.

The network loss function is

$$L_h(x, c, l, g) = \frac{1}{N}(L_{h\_c}(x, c) + \alpha L_{h\_l}(x, l, g)) \tag{2}$$

The loss function includes two parts which are confidence loss and location loss. Where N represents the number of default boxes for matching to groundtruth boxes. $x_{ij} = 0, 1$ indicates that the ith default box matches the jth groundtruth boxes. C indicates confidence. l indicates default box and g indicates groundtruth box. L_{h_c} means confidence loss, it uses softmax loss, while L_{h_l} means location loss, it uses smooth $L1$ loss. Parameter α is utilized to adjust the ratio between confidence loss and location loss and default $\alpha = 1$.

3.2. Feature fusion module

The feature fusion module in Figure 6 borrows the idea of deconvolution in DSSD network. And its structure is shown as Figure 7.

The specific method is to deconvolve the feature map of the Conv6_2 layer twice to get Deconv6_2 layer so that the Deconv6_2 layers size is the same as Conv4_3 layer. Then a 3×3 convolutional layer is used after Conv4_3 layer and Deconv6_2 layer. It can learn better features for fusion. After that, a regular layer with different scales such as 10 and 20 is utilized behind these two layers. Finally, the feature maps of these two layers are merged. According to reference

[18], the element-wise product which is a kind of fusion method can get the highest accuracy, so this method is chosen as the fusion method. As a result, the new layer Fusion_conv43_conv62 is used to detect small targets.

Its function is

$$M_F = RELU(M_S \cdot M_D) \tag{3}$$

where $RELU$ is an activation function. Its expression is

$$RELU(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{4}$$

M_F is the feature map matrix generated after the feature fusion of Conv4_3 layer and Conv6_2 layer (the green block in Figure 7. M_S is the feature map matrix before element-wise product for shallow layer Conv4_3. M_D is the feature map matrix before element-wise product for deep layer Conv6_2. And where

$$M_S = Norm[Conv(M_{4\_3})] \tag{5}$$

$$M_D = Norm_{10}\left\{Conv[Deconv(M_{6\_2})]\right\} \tag{6}$$

where $Norm$ is batch normalization [24]. M_{4_3} is the feature map matrix of Conv4_3 layer, and M_{6_2} is the feature map matrix of Conv6_2 layer.

4. Comparative experimental results and discussion

4.1. Hand databases

In order to recognize the above SHRI hand gestures, a set of Space Robot Simple Sign Language (SRSSL) databases is made. The hand database collects hand RGB images from 6 volunteers. Each person's hand gestures include the 8 kinds of SHRI hand gestures, and they are split into 5 different sizes. Each person has 1000 hand images, so the total number of images is 6000. Five peoples hand images (5000 images) are chosen as train data, and the other ones

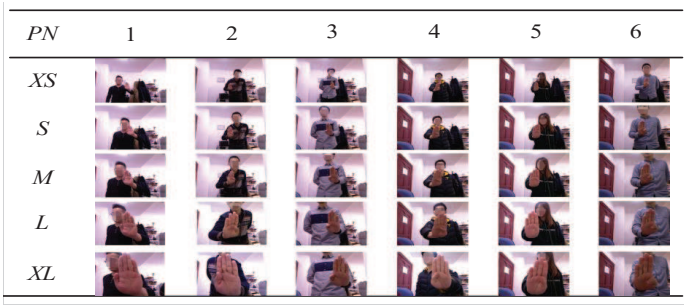

Figure 8: SRSSL database.

hand images (1000 images) used as test data. Parts of the SRSSL database ("Begin to control" hand gesture images) are shown in Figure 8. Where PN represents person number, XS, S, M, L and XL represent five different sized hand images which are extra small size, small size, medium size, large size and extra large size.

Because the hand database is small, training directly on it cannot get a good result. Therefore, combining this database with the existing hand detection database can get better results. There are two public well-known hand databases used for detection and location. One is Oxford hand database [25] made by Oxford University, and the other one is Egohands database [26] made by Indiana University.

4.2. Network training framework

The network training framework consists of two steps. At the beginning, the structure of SSD300 is changed to FF-SSD300 architecture. In order to greatly reduce the training time, the transfer learning [27] method is used in this step. We train the existing model SSD300_VOC0712 on Egohands and Oxford hand databases, and get a FF-SSD300_HandNet. After that, retrain this network on SRSSL database and get a FF-SSD300_SpacehandNet at last. Thus, the FF-SSD300_SpacehandNet can detect and locate different space hand gestures. The specific experimental framework is shown in Figure 9.

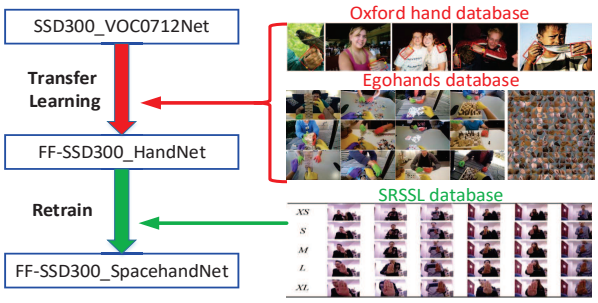

Figure 9: Experiment framework.

4.3. Experiment results

The experiments are conducted in Caffe. The SSD network, the DSSD network, and the FF-SSD network are trained and tested according to the network training method of Figure 9 respectively. During the training, use stochastic gradient descent (SGD), and the start learning rate is 0.001, momentum is 0.9 and weight decay is 0.9. The changing mode learning rate choose multistep mode, and the learning rate successively drops by 10 at 10000, 20000 and 25000 iterations. And performed a test after each 3000 training. The entire training process is conducted under Intel Core i5-6400 CPU, NVIDIA GeForce GTX 1060 6GB GDDR5, 16GB RAM. And the software is selected Caffe in the Ubuntu 16.04 64bit OS system.

4.3.1. mAPS of different databases

Two well-known public hand databases are used for hand detection and localization. One is Oxford hand database, and the other one is Egohands database. The Oxford hand database is collected from various different public image dataset source such as Skin Dataset [28] and PASCAL VOC 2007 and 2010. Most of its images are random hand gestures in our daily life. Using this database only in the first training step cannot get a good result. Then the Egohands database is used, it contains 48 different videos of egocentric interactions with pixel-level ground-truth annotations. Most of its images are hand gestures that interact with objects. And it can get a better result than

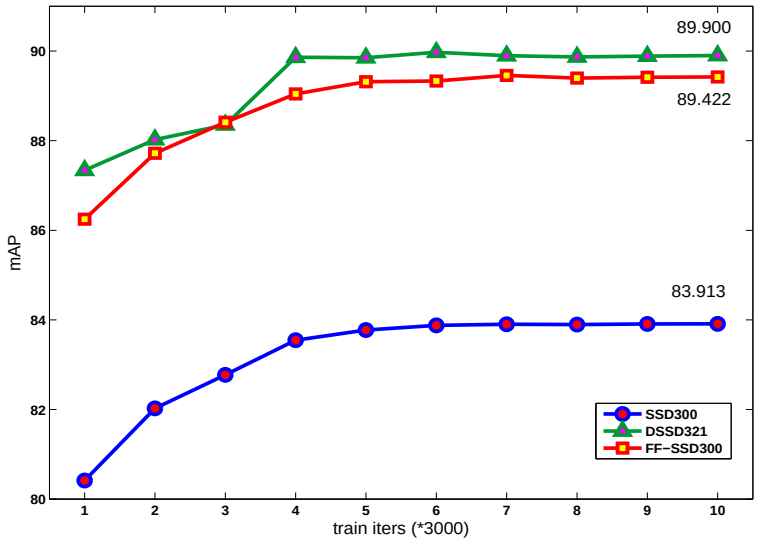

Figure 10: Test mAPs of the three networks.

Oxford hand database. At last, these three hand databases are combined and we get a best result on FF-SSD300 network. The specific accuracies of these databases are shown as Table 2.

Table 2: mAPs in different databases.

Database	mAP
Egohands + SRSSL	83.9
Oxford hand + SRSSL	73.7
Egohands + Oxford hand + SRSSL	89.4

4.3.2. mAPs of three networks

The test mAP curves for each network trained under SRSSL database are shown in Figure 10.

As can be observed in Figure 10, after training, the final mAPs of the SSD300, DSSD321, and the FF-SSD300 on the SRSSL test set reached 83.913,

89.900, and 89.422, respectively. The accuracy of the FF-SSD (89.422) is much higher than that of the SSD (83.913). This is because there are a large number of small-sized hand images in the test data, and the FF-SSD network has higher detection accuracy for small-sized hand images than SSD networks. While the accuracy of the FF-SSD network (89.422) is slightly lower than that of the DSSD network (89.900). This is because on the one hand, the input image size of DSSD is slightly bigger than that of the FF-SSD, since there is no DSSD300 network, we have to use the DSSD321 network. On the other hand, DSSD merges multiple deep layers with multiple shallow layers. The effect of information extraction after feature fusion is theoretically superior to that of the FF-SSD network with only one layer of feature fusion.

4.3.3. Speeds of three networks

However, this project is required to analyze the network in terms of not only accuracy but also speed. Therefore, the mAPs and FPS of the three networks are tested separately. The results are presented in Figure 11.

It can be seen from Figure 11 that although the accuracy of DSSD321 is highest among these three networks, but its speed is very low and cannot satisfy the necessary of real-time system. While the SSD300 can get a fastest speed among them, but the accuracy of it is not sufficient. The improved SSD300 cannot only get a high accuracy but also have a fast speed, so it is most suitable in the hand-based SHRI system.

4.3.4. Different hand sizes

The test results of the three networks on SRSSL hand images of different sizes are shown in Table 3.

It can be seen from Table 3 that the accuracies of DSSD321 and FF-SSD300 are similar in every sized hand images. When the hand sizes are big enough such as L and XL, these three network can get similar accuracies. And when hand size is small such as XS and S, the SSD300 network cannot get good results, while the DSSD321 and FF-SSD300 can get better results than that of

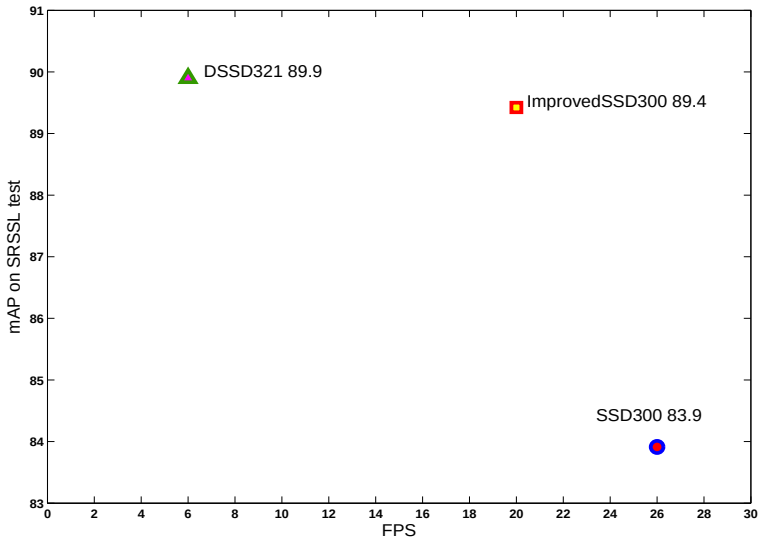

Figure 11: Speeds and accuracies of the three networks.

Table 3: mAPs on different sized hand images.

Deep learning models	SSD300	DSSD321	FF-SSD300
XS	62.38	88.47	86.34
S	83.75	87.74	87.83
M	89.43	88.53	89.96
L	90.50	92.32	90.34
XL	93.48	92.43	92.21

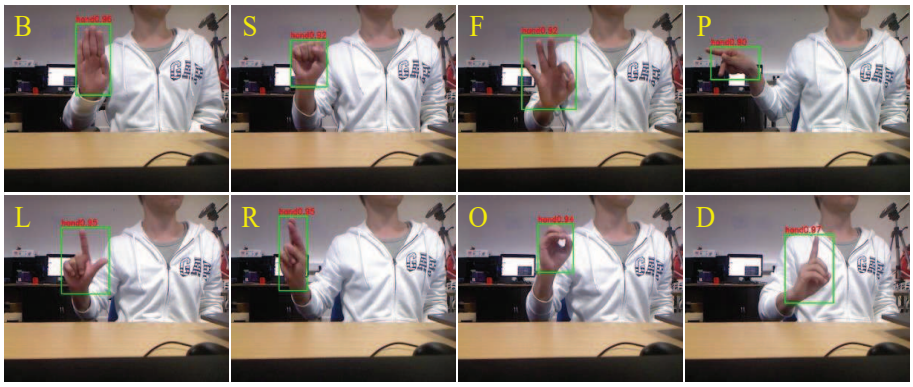

Figure 12: Hand detection and location on different hand gestures.

SSD300. This can prove that the FF-SSD can get expert performs when hand size is small.

4.3.5. Different hand gestures

In order to prove that the FF-SSD network can detect different hand gestures in SHRI hand gestures dataset, different hand gestures are tested on the FF-SSD network. The results are shown in Table 4.

Table 4: mAPs on different hand gestures.

Hand gesture	B	S	F	P	L	R	O	D
mAP	90.76	88.01	88.38	89.42	89.64	89.73	86.96	92.00

It can be seen from Table 4 that the detection accuracies of all hand gestures are above 88%. The *S* hand gesture gets the smallest accuracy which is 88.01%, and the *D* hand gesture gets the biggest accuracy which is 92.0%. So, the FF-SSD network can get expert performs on all different hand gestures. And some of the hand detection and localization images are shown in Figure 12, and the video of hand detection is available at *https://www.youtube.com/watch?v=mG79nr1APZ4*.

5. Conclusion remakes and future directions

A novel feature-map-fused SSD network for hand detection and localization is proposed in this paper, and it's used in the astronaut-AAR hand-gesture interaction. The main contributions of the paper are:

- Our FF-SSD method can get higher accuracies on both complicated hand gestures and small-sized hand images.
- Our method can balance both speed and accuracy with a satisfactory.
- Our method has been successfully implemented in the astronaut-AAR interaction system.

In addition, experiments on the FF-SSD network have mainly the following contributions:

- Homemade a SRSSL database for astronaut's hand detection and localization, and use transfer learning to combine the SRSSL database with two public hand database.
- Compare the results of FF-SSD with the results of SSD and DSSD, the experimental results can prove that the FF-SSD network cannot only increase the detection accuracy of pony-size hands but also get a real-time speed.
- By testing the FF-SSD network on images which have different hand gestures, the results can prove that the FF-SSD can detect and locate the SHRI hand gestures effectively and efficiently.

The FF-SSD network can get a good performance for hand detection and localization, but there are also some defects, for example, when two hands are very close with each other, the method will detect them as one hand. And this method cannot distinguish right hand and left hand. So, it's very important to deal with these problems in the future. In addition, the FF-SSD method will be favorable to add in hand gesture recognition.

Acknowledgments

The authors would like to acknowledge the support from the Research Fund of China Manned Space Engineering (050102), the Key Research Program of the Chinese Academy of Sciences (Y4A3210301), the Natural Science Foundation of China under Grant No. 51775541, 51575412, 51575338 and 51575407, the EU Seventh Framework Programme (FP7)−ICT under Grant No. 611391, and the Research Project of State Key Lab of Digital Manufacturing Equipment and Technology of China under Grant No. DMETKF2017003.

References

[1] G. A. Dorais, Y. Gawdiak, The personal satellite assistant: an internal spacecraft autonomous mobile monitor, in: Aerospace Conference, 2003. Proceedings. 2003 IEEE, Vol. 1, IEEE, 2003, pp. 1–348.

[2] T. Fong, M. J. Micire, T. Morse, E. Park, C. Provencher, V. To, D. Wheeler, D. Mittman, R. J. Torres, E. Smith, Smart spheres: a telerobotic free-flyer for intravehicular activities in space.

[3] M. Bualat, J. Barlow, T. Fong, C. Provencher, T. Smith, Astrobee: Developing a free-flying robot for the international space station, in: AIAA SPACE 2015 Conference and Exposition, 2015, p. 4643.

[4] T. Smith, J. Barlow, M. Bualat, T. Fong, C. Provencher, H. Sanchez, E. Smith, Astrobee: A new platform for free-flying robotics on the international space station.

[5] T. I. Murphy, Hello, i am cimon!, `http://www.airbus.com/newsroom/press-releases/en/2018/02/hello--i-am-cimon-.html`, accessed February, 2018.

[6] J. Liu, Q. Gao, Z. Liu, Y. Li, Attitude control for astronaut assisted robot in the space station, International Journal of Control, Automation and Systems 14 (4) (2016) 1082–1095.

[7] Q. Gao, J. Liu, T. Tian, Y. Li, Free-flying dynamics and control of an astronaut assistant robot based on fuzzy sliding mode algorithm, Acta Astronautica 138 (2017) 462–474.

[8] Q. Gao, J. Liu, Z. Ju, Y. Li, T. Zhang, L. Zhang, Static hand gesture recognition with parallel cnns for space human-robot interaction, in: International Conference on Intelligent Robotics and Applications, Springer, 2017, pp. 462–473.

[9] W. He, Y. Dong, Adaptive fuzzy neural network control for a constrained robot using impedance learning, IEEE transactions on neural networks and learning systems 29 (4) (2018) 1174–1186.

[10] W. He, S. S. Ge, Cooperative control of a nonuniform gantry crane with constrained tension, Automatica 66 (2016) 146–154.

[11] W. He, T. Meng, X. He, S. S. Ge, Unified iterative learning control for flexible structures with input constraints, Automatica 96 (2018) 326–336.

[12] W. He, Z. Li, C. P. Chen, A survey of human-centered intelligent robots: issues and challenges, IEEE/CAA Journal of Automatica Sinica 4 (4) (2017) 602–609.

[13] K. Lenc, A. Vedaldi, R-cnn minus r, arXiv preprint arXiv:1506.06981.

[14] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[15] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[16] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.

[18] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, Dssd: Deconvolutional single shot detector, arXiv preprint arXiv:1701.06659.

[19] Z. Li, F. Zhou, Fssd: Feature fusion single shot multibox detector, arXiv preprint arXiv:1712.00960.

[20] L. Zheng, C. Fu, Y. Zhao, Extend the shallow part of single shot multibox detector via convolutional neural network, arXiv preprint arXiv:1801.05918.

[21] J. Liu, Y. Luo, Z. Ju, An interactive astronaut-robot system with gesture control, Computational intelligence and neuroscience 2016.

[22] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, J. Wu, Feature-fused ssd: fast detection for small objects, in: Ninth International Conference on Graphic and Image Processing (ICGIP 2017), Vol. 10615, International Society for Optics and Photonics, 2018, p. 106151E.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[24] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

[25] A. Mittal, A. Zisserman, P. H. Torr, Hand detection using multiple proposals., in: BMVC, Citeseer, 2011, pp. 1–11.

[26] S. Bambach, S. Lee, D. J. Crandall, C. Yu, Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1949–1957.

[27] K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big Data 3 (1) (2016) 9.

[28] M. J. Jones, J. M. Rehg, Statistical color models with application to skin detection, International Journal of Computer Vision 46 (1) (2002) 81–96.